Поиск музыки в социальной сети по запросу с орфографическими ошибками

Ерзикова Юлия
Галкина Алёна
Филатова Анастасия

Райский Андрей
Перваков Григорий
Зархидзе Никита

Группа 341

Информационный поиск, СПБГУ, кафедра ИАС, весна 2018 г.

Постановка задачи

Поиск музыкальных произведений по запросам с орфографическими ошибками и опечатками

Сбор данных

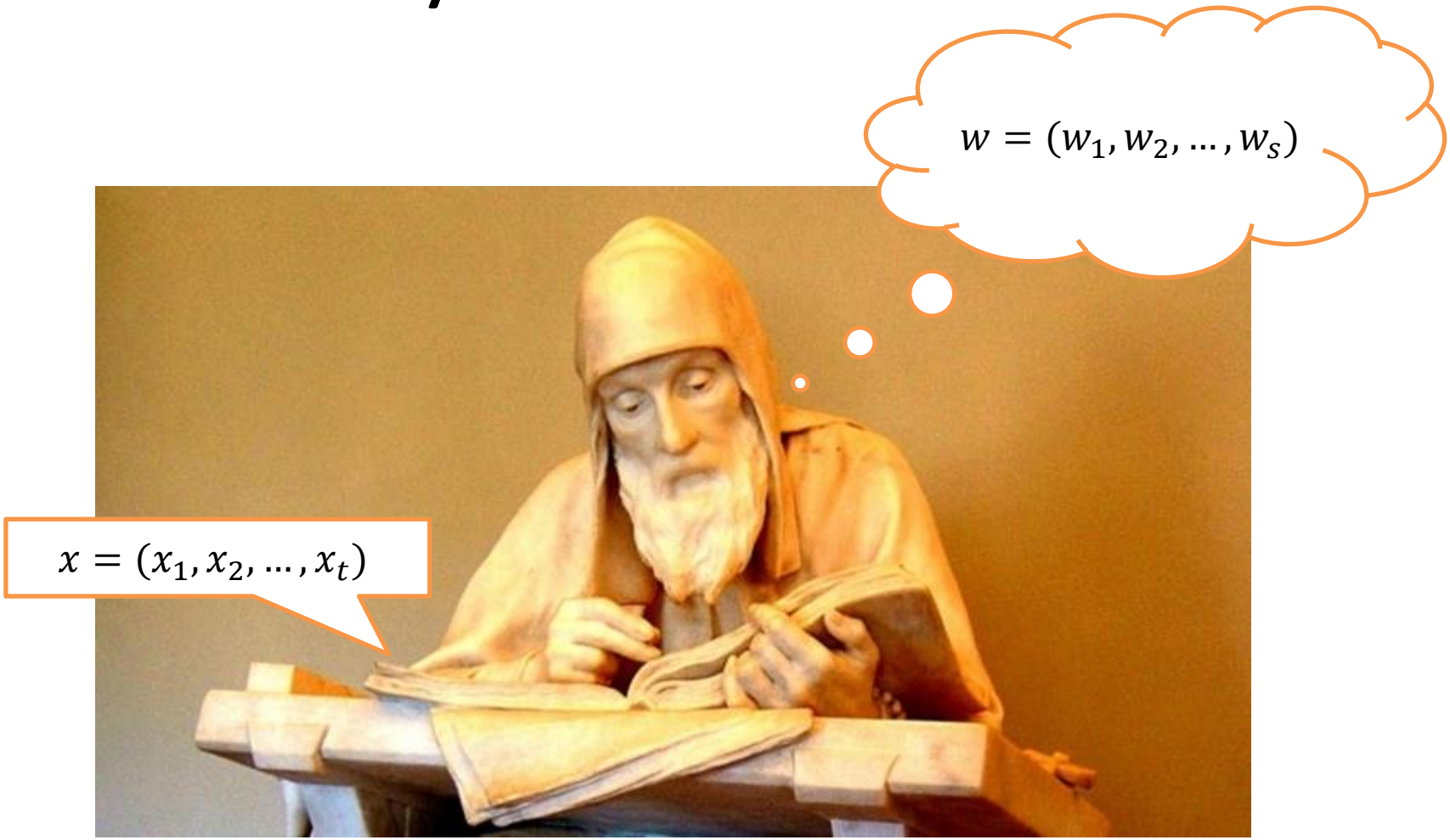
- База данных наиболее запрашиваемых англоязычных песен с сайта lyrics.net (на данный момент lyrics.com)

wc_lyricsnet_albums	
id	PK
lyricsnet_id	
artist_id	
name	
year	

wc_lyricsnet_artists	
id	PK
lyricsnet_id	
lyricsnet_tag	
name	
albums_count	
songs_count	
parsed	

wc_lyricsnet_songs	
id	PK
lyricsnet_id	
artist_id	
album_id	
title	
parsed	
views_count	
view_last_time	
skipped	
archived	

Noisy channel model



$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} P(w|x) = \underset{w \in V}{\operatorname{argmax}} \frac{P(x|w)P(w)}{P(x)}$$

$$\boxed{\hat{w} = \underset{w \in V}{\operatorname{argmax}} P(x|w)P(w)}$$

- $P(x|w)$ $-$ $likelihood(channel\ model)$
- $P(w) - prior\ probability$

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} \log P(w) + \log P(x|w)$$

$$P(x|w) = P(f|w \rightarrow x) = P(f) = \prod_{i=1}^{k} P(f_i)$$

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} \log P(w) + \sum_{i} \log P(f_i)$$

$$P(x|w) = \begin{cases} \dfrac{\text{del}[x_{i-1}, w_i]}{\text{count}[x_{i-1} w_i]}, \text{ if deletion} \\ \dfrac{\text{ins}[x_{i-1}, w_i]}{\text{count}[w_{i-1}]}, \text{ if insertion} \\ \dfrac{\text{sub}[x_i, w_i]}{\text{count}[w_i]}, \text{ if substitution} \\ \dfrac{\text{trans}[w_i, w_{i+1}]}{\text{count}[w_i w_{i+1}]}, \text{ if transposition} \end{cases}$$

Модифицированный алгоритм Дамерау-Левенштейна

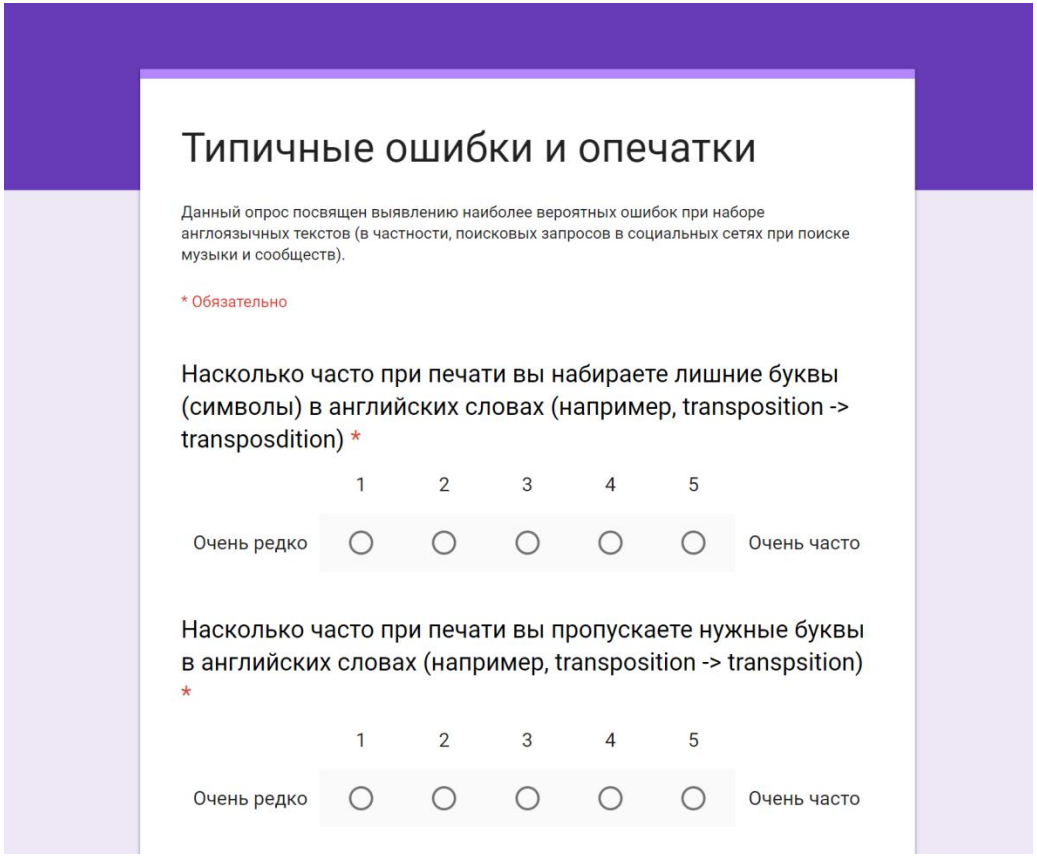
Базовые операции:

- ADD
- DELETE
- SUBSTITUTE
- TRANSPOSE

Дополнительные операции:

- DOUBLE
- SINGLE

Сбор статистики об ошибках реальных пользователей



Типичные ошибки и опечатки

Данный опрос посвящен выявлению наиболее вероятных ошибок при наборе англоязычных текстов (в частности, поисковых запросов в социальных сетях при поиске музыки и сообществ).

* Обязательно

Насколько часто при печати вы набираете лишние буквы (символы) в английских словах (например, transposition -> transposdition) *

	1	2	3	4	5	
Очень редко	○	○	○	○	○	Очень часто

Насколько часто при печати вы пропускаете нужные буквы в английских словах (например, transposition -> transpsition) *

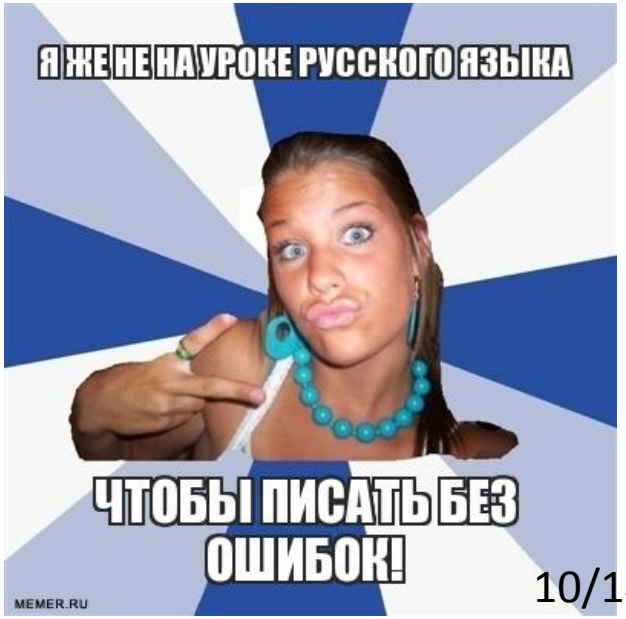
	1	2	3	4	5	
Очень редко	○	○	○	○	○	Очень часто

add	delete	transpose	single/double	substitute
2,90	2,90	2,99	2,62	2,83

Немного статистики

Согласно лингвистическим исследованиям:

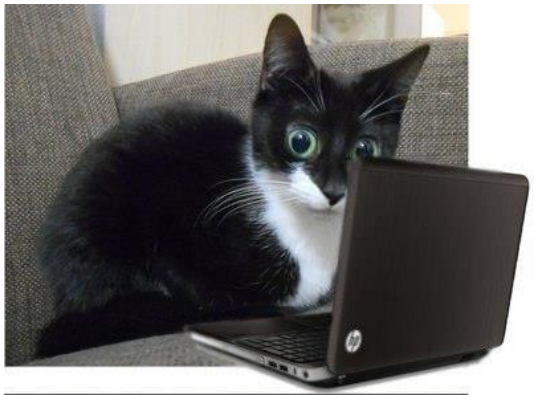
- 80% всех ошибочных написаний слова находятся на редакционном расстоянии, равном одному от его верного написания;
- Почти в 100% случаев редакционное расстояние равно двум

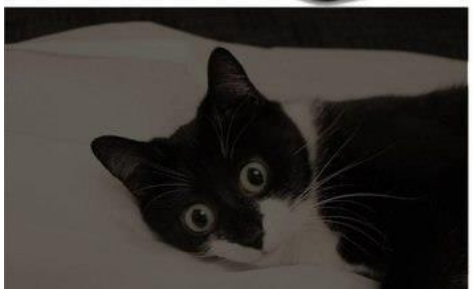


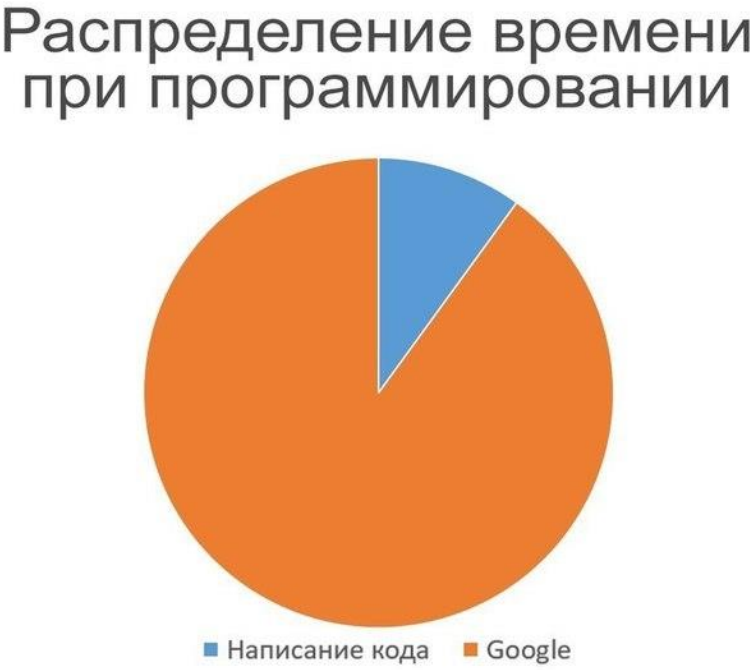
P.S. Некоторые индивиды, конечно, ошибаются чаще ;)

Этапы обработки запросов:

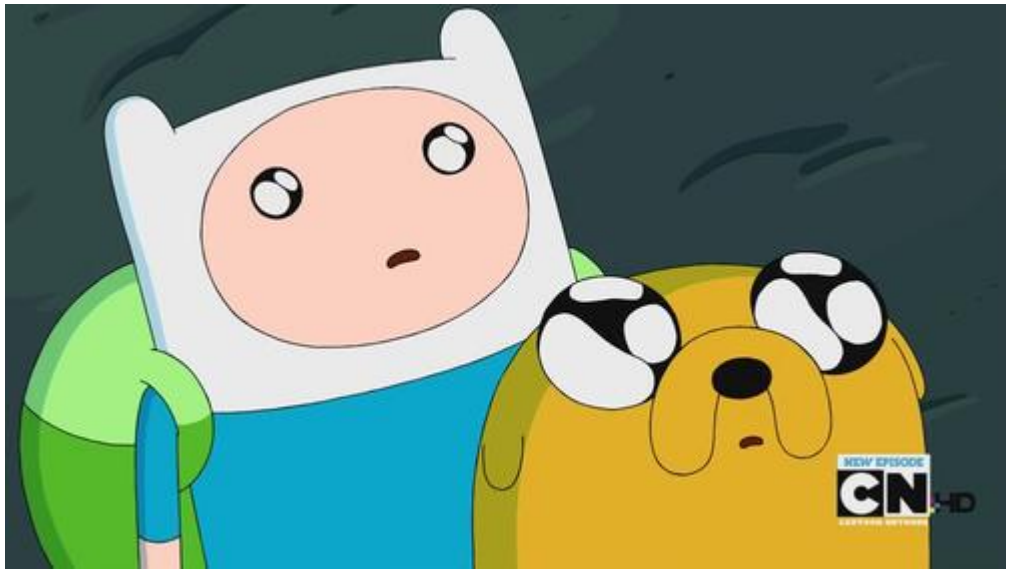
- Разбиваем поисковый запрос на слова
- Для каждого слова применяем алгоритм Noisy Channel и находим список основных песен-кандидатов
- Пересекаем списки для всех слов в запросе, получаем суженный набор кандидатов
- Выбираем из данного набора оптимальных кандидатов, используя метрику Левенштейна







It’s demonstration time!



Источники

- Daniel Jurafsky & James H. Martin «Speech and language processing»
- Kernighan, M. D., Church, K.W., and Gale,W. A. (1990). A spelling correction program base on a noisy channel model.
- Norvig, P. (2009). Natural language corpus data (http://norvig.com/ngrams/)

Спасибо за внимание!